



Development and clinical impact assessment of a machine-learning model for early prediction of late-onset sepsis

Merel (A.M.) van den Berg^{a,1}, O'Jay (O.A.G.) Medina^{b,1}, Ingmar (I.P.) Loohuis^b,
 Michiel (M.) van der Flier^c, Jeroen (J.) Dudink^a, Manon (M.J.N.L.) Benders^a,
 Richard (R.T.) Bartels^b, Daniel (D.C.) Vijlbrief^{a,*}

^a Department of Neonatology, Wilhelmina Children's Hospital, UMC Utrecht, Utrecht, the Netherlands

^b Department of Digital Health, UMC Utrecht, Utrecht, the Netherlands

^c Department of Pediatric Infectious Disease, Wilhelmina Children's Hospital, UMC Utrecht, Utrecht, the Netherlands

ARTICLE INFO

Keywords:

Preterm infants
 Late-onset sepsis
 Machine learning
 Algorithm
 NICU
 Early-warning
 Impact assessment

ABSTRACT

Background and aim: Preterm infants are prone to neonatal infections such as late-onset sepsis (LOS). The consequences of LOS can be severe and potentially life-threatening. Unfortunately, LOS often presents with unspecific symptoms, and early screening laboratory tests have limited diagnostic value and are often late. This study aimed to build a predictive algorithm to aid doctors in the early detection of LOS in very preterm infants.

Methods: In a retrospective cohort study, all consecutively admitted preterm infants (GA \leq 32 weeks) from 2008 until 2019 were included. They were classified as LOS or control according to blood culture results, currently the gold standard. To generate features, routine and continuously measured oxygen saturation and heart rate data with a minute-by-minute sampling rate were extracted from electronic medical records. Care was taken not to include variables indicative of existing LOS suspicion. The timing of a positive blood culture served as a proxy for LOS-onset. An equivalent timestamp was generated in gestational-age-matched control patients without a positive blood culture. Three machine learning (ML) techniques (generalized additive models, logistic regression, and XGBoost) were used to build a classification algorithm. To simulate the performance of the algorithm in clinical practice, a simulation using multiple alarm thresholds was performed on hourly predictions for the total hospitalization period.

Results: 292 infants with LOS were matched to 1497 controls. The median gestational age before matching was 28.1 and 30.3 weeks, respectively. Evaluation of the overall discriminative power of the LR algorithm yielded an AUC of 0.73 ($p < 0.05$) at the moment of clinical suspicion ($t = 0$). In the longitudinal simulation, our algorithm detects LOS in at least 47% of the patients before clinical suspicion without exceeding the alarm fatigue threshold of 3 alarms per day. Furthermore, medical experts evaluated the algorithm as clinically relevant regarding the feature contributions in the model explanations.

Conclusions: An ML algorithm was trained for the early detection of LOS. Performance was evaluated on both prediction horizons and in a clinical impact simulation. To the best of our knowledge, our assessment of clinical impact with a retrospective simulation on longitudinal data is the most extensive in the literature on LOS prediction to date. The clinically relevant algorithm, based on routinely collected data, can potentially accelerate clinical decisions in the early detection of LOS, even with limited inputs.

1. Introduction

Neonatal sepsis is a severe infectious disease and a significant cause of neonatal morbidity and mortality worldwide [1]. Particularly,

preterm infants and very-low-birth-weight infants (VLBW) are prone to neonatal sepsis [2–4]. Neonatal sepsis is often categorized based on the age of onset. Early onset sepsis (EOS) occurs in the first days of life and is usually caused by bacteria which are transmitted from mother to child

* Corresponding author. Department of Neonatology, Lundlaan 6, 3584 EA, Utrecht, the Netherlands.

E-mail address: D.C.Vijlbrief@umcutrecht.nl (D.(D.C.) Vijlbrief).

¹ Both authors contributed equally.

during labor or in the prenatal period. Late onset sepsis (LOS) occurs after the first week of life and can be caused by a variety of bacteria which are usually acquired from the environment of the child [5]. This article focuses on late onset sepsis. The consequences of late-onset sepsis (LOS) include neurological impairment, bronchopulmonary dysplasia, prolonged hospitalization, multi-organ system failure, and death [2,3,6]. Globally, neonatal sepsis is the third highest cause of neonatal deaths, accounting for 13%. In high-income countries, 39% of all cases of neonatal sepsis lead to major disability or even death, despite initiation of treatment. Mortality rates range from 5% to 20% [7].

LOS is difficult to timely recognize as it often presents with unspecific symptoms. Patients often demonstrate subtle signs of sickness or ambiguous clinical signs which can easily be misunderstood for other neonatal disease [6]. Next, screening laboratory tests have limited diagnostic value [2]. A positive blood culture is the gold standard for diagnosing LOS as it is currently the only method to demonstrate presence of a bacterium in the blood stream of a patient. However, this test has an inherent delay due to a natural growth period up until multiple days. Also, specificity is limited in the light of antibiotic treatment and small specimen volumes [2]. Therefore, a clinical diagnosis is difficult to ascertain and often late [8]. As repercussions can be immense, early detection and treatment are crucial to the outcome. Therefore, clinicians start empirical treatment with broad-spectrum antibiotics before an official diagnosis is confirmed by blood culture, meaning newborns with clinical signs of LOS receive antibiotic treatment as soon as suspicion arises. Unfortunately, this diligence also results in inappropriate treatment as also true culture-negative patients receive antibiotic treatment. Consequently, there is antibiotic overtreatment within the NICU [2,4,6]. The adverse effects of prolonged empirical antibiotic treatment in pre-term infants are not without harm. These include antibiotic resistance, fungal infections, necrotizing enterocolitis (NEC), and death [9]. Additionally, growing evidence indicates that antibiotics increase the risk of obesity and asthma later in life [2].

As LOS is difficult to diagnose timely and precisely, it becomes clear a reliable screening tool for LOS is urgently warranted. Several studies have already contributed by developing machine learning algorithms. These algorithms are often based on demographic data, vital signs, and laboratory data. Objectives of these algorithms vary from assisting the bedside clinician in treatment decisions on the day of clinical suspicion to an accelerated trigger indicating sepsis in children in the NICU. However, clinical implementation is limited, and some algorithms are potentially biased by including variables that hold prior knowledge. For example, this is the case when including laboratory measurements in the algorithm, as a clinician's decision of drawing blood already holds information about a suspicion of disease [2,8,10–12].

This study aimed to create a clinically relevant machine learning algorithm to be used as a tool for individual, earlier than suspected detection of LOS in a group of very preterm infants. In order to both raise doctor's attention to potential LOS patients at an earlier stage than presentation of symptoms, and to enable implementation across multiple hospitals, the algorithm was developed exclusively on routinely collected data and vital signs. Next to validation in a cross-sectional dataset, algorithm performance has also been validated in a more realistic longitudinal simulation, which is unique in the field.

2. Materials and methods

The institutional Medical Research Ethics Committee (MREC) approved this research of the University Medical Centre Utrecht (MREC number 17/894). They permitted the exemption of asking for informed consent at the UMC Utrecht based on the study's retrospective nature, the large number of patients, and the use of pseudo-anonymized data.

2.1. Study design and patients

Retrospective data was reviewed from a large cohort of very preterm

infants hospitalized at the NICU of the Wilhelmina Children's Hospital (WKZ) between April 2008 and May 2019, regardless of their admission status. The NICU of the WKZ is one of the largest tertiary referral centres in the Netherlands and annually admits 600 patients, of which 400 are prematurely born. Included infants had a gestational age of ≤ 32 weeks, were admitted to the hospital within 48 h after birth and had complete medical records from admission until at least 30 days after birth or until discharge. Infants with severe congenital syndromes or those who died in the first 96 h were excluded from the study.

Variable selection was based on literature review and clinical expertise. Baseline demographics were collected, such as gestational age, sex, hospital of birth, birth weight, physical appearance (skin colour), admission, and discharge times. Also, low-frequency (at most one sample per minute) vital parameters such as temperature, blood pressure, respiratory rate, oxygen saturation, heart rate, and laboratory data like CRP, white blood cell count, neutrophil count, and thrombocyte count were assessed. In addition, blood-culture specimen data was collected. Finally, data were collected on nutrition and the presence of central venous lines, arterial and venous umbilical lines and PICC lines.

The goal was to develop a predictive model that could help doctors detect the risk of sepsis at an earlier stage than current clinical suspicion. Therefore, it was essential to select variables that do not implicitly carry information about a clinical sepsis suspicion through the clinical behaviour [13]. For example, C-reactive protein (CRP) is often measured upon suspicion of an infection. As a result, after screening the data generation process, only a small set of the collected variables were eventually included in the modelling stage.

2.2. Patient selection

Patients were defined as LOS patients when a positive blood culture was drawn within the window of 72 h until 30 days after birth. The lower bound was set due to a pragmatic division between early-onset sepsis (EOS) and LOS, even though a formal differentiation between the two does not exist. The onset of LOS was defined as the time a blood culture was drawn, as registered in the patient records. This serves as a proxy for the first clinical suspicion ($t = 0$). When the latter demonstrated a positive result, LOS was confirmed, and the patient was included in the LOS group. Patients with a positive blood culture outside the pre-set time window or without a blood culture were included in the control group. In addition, patients with a negative blood culture within the time window were included in the control group, as sepsis diagnosis was not definitive for these patients (Fig. 1). Furthermore, in these infants, there was only a clinical suspicion of sepsis during a short period of their admission period. The rest of the data gathered was equivalent

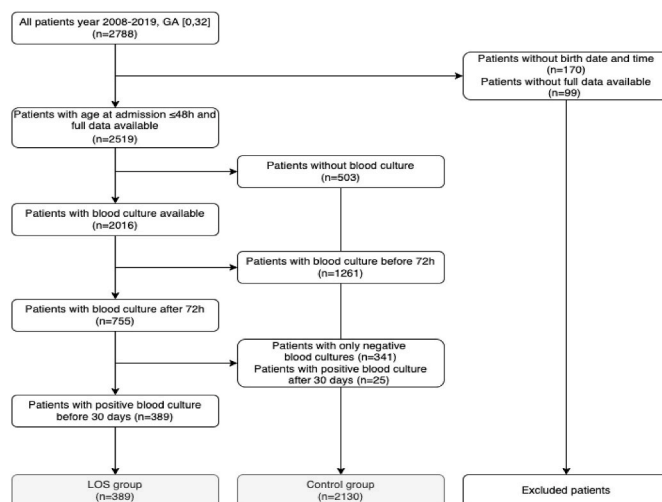


Fig. 1. Patient inclusion flowchart.

to the infants without a blood culture and of potential value for the model building.

For model building purposes, control patients were matched with LOS patients to create a cross-sectional dataset. Specifically, for each LOS patient, at least one control patient was selected with a similar gestational age (± 2 days) and at the same time since birth in hours. E.g., for a LOS patient with a positive blood culture 100 h after birth, a control patient data 100 h after birth was used. Important to note here is that for control patients, multiple data points can be sampled. Control patients were restrained to be included only once per 24 h, dropping samples that were within 24 h of an earlier match. The proposed matching procedure ensures a degree of similarity between the LOS and control groups since patients of similar gestational and postmenstrual age are included. Moreover, this method also partly counters the effect of class imbalance. However, it does not entirely remove infants' maturational effects since more than one control patient can be matched to each LOS patient, differing from case-control studies in that respect.

2.3. Descriptive statistics

For a description of the patients, data was explored separately for the two patient groups (LOS and control) and split into training and test data (control-train, control-test, LOS-train, LOS-test). Importantly, the results of this exploratory analysis were derived from the raw data, thus before further processing or feature engineering. Demographic and clinical characteristics are presented as means and standard deviations in the case of normal distributions and medians and interquartile ranges in the case of non-normal distributions. For testing normality, the Shapiro-Wilk test was used. Differences for continuous variables were assessed using the student's t-test for normal distributions, the Mann-Whitney *U* test for non-normally distributed samples, and the chi-square test for categorical variables. P-values below 0.05 were considered statistically significant.

2.4. Data processing

The scheme for model building, evaluation and testing is shown in Fig. 2. To validate the algorithm, patients were randomly assigned to the training set (75%) for development or the test set (25%) for final evaluation. First, observational data was processed which included demographic data like sex, gestational age and specimen time stamps. Here, all patients were categorized as either LOS or control, based on the result and timing of their blood culture specimen. A fraction of the

patients had blood culture specimen times at exactly midnight, which is likely an artifact in the data and considered illogical by medical experts. These contaminated timestamps were imputed to noon, which is the most frequent time for blood cultures. Minute-by-minute vital parameter data was combined per patient and anomalous values were filtered to prevent the effects of extreme outliers (allowed values: $1 \leq HR \leq 275$ beats per minute (bpm); $SpO_2 \leq 100\%$) (see supplementary material Sect. 1.1 for details). Features were generated using the timestamped electronic medical records (EMR) over predefined 4-h intervals to monitor acute trends. Features were computed during this interval through simple statistics (mean, minimum and variance) and more complex clinically relevant features. The latter being counts of bradycardia ($HR < 85$ bpm), tachycardia ($HR > 180$ bpm) and saturation drops ($SpO_2 < 85\%$). A minimum of 90% completeness in each vital parameter in the 4-h interval was required for feature computation. Missing values were set to NaN and ignored in the computation of the features.

The time of the blood culture served as a proxy for LOS onset ($t = 0$). For LOS patients, only one data point was used for model building, corresponding to the first LOS episode. For the control patients, the time of blood culture was fabricated based on the matched LOS patient (see Sect. 2.2).

2.5. Building and evaluating the algorithm

2.5.1. Model training

During model development, three criteria were explicitly considered next to the overall performance. First, the model should contain features that can detect actionable insights, meaning that features should be able to capture acute changes in patient well-being. Secondly, the underlying logic should be interpretable for clinicians. Three different types of models were assessed, namely, logistic regression (LR), generalized additive models (GAMs), and eXtreme Gradient Boosting (XGBoost) [14]. These models were chosen since they are representative of an interpretable linear (LR) and non-linear model (GAM) and a state-of-the-art black-box tree-based classifier (XGBoost). Interpretable models are easier to understand by clinicians compared to more complex black-box models. The model decisions should at least be considered well explainable (i.e., model coefficients or SHAP values [15]). Finally, the model features were thoroughly examined for their potential use in clinical practice, meaning that features should not be dependent on pre-existing clinical suspicion.

Initially, nested cross validation was performed to select an appropriate hyperparameter grid for the different classifiers (Fig. 2, for details see supplementary material Sect. 1.3). Then, the different classifiers were evaluated using grouped cross-validation on the cross-sectional data. In cross-validation, the data was separated in *n* folds. Each time *n*-1 folds were used to train a model while the remaining fold was used for evaluation. Grouping was done on a patient level and stratified using scikit-learn its *StratifiedGroupKFold* method, meaning that within one iteration patients in an evaluation fold were never present in the corresponding training folds and labels were distributed approximately evenly between training and evaluation data (Fig. 2). Grid search was used to determine optimal hyperparameters. These were then frozen and the grouped cross-validation procedure repeated to compute performance metrics. The discriminative power of the classifiers was assessed using the area-under-the-curve (AUC). 95% confidence intervals (C.I.) were computed using 1000-fold bootstrapping and AUCs between models were compared using DeLong's test [16]. Eventually, the overall performance was evaluated based on the aforementioned AUC (at $t = 0$), a preliminary impact simulation (see Sect. 2.5.2), and the level of interpretability. The classifiers, including the corresponding optimal set of hyperparameters, were retrained on the entire training dataset and evaluated on the test set.

2.5.2. Preliminary impact assessment

In the case of an early warning system, metrics such as the AUC at $t =$

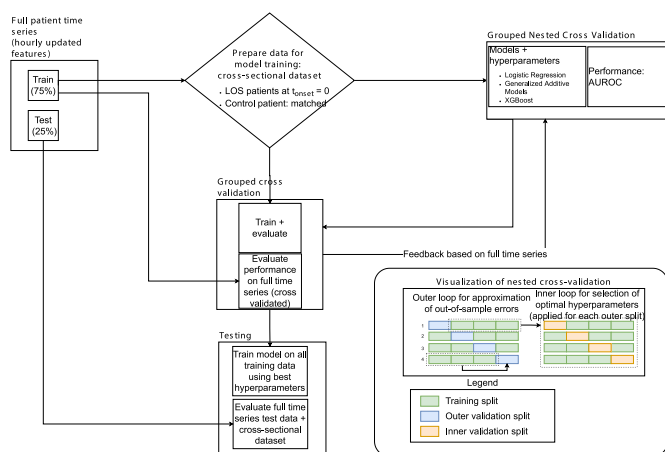


Fig. 2. Scheme for model building, evaluation and testing. The early-warning algorithm was trained on a cross-sectional dataset. Additional evaluation to test real-life performance was done on the full longitudinal dataset. The (nested) cross-validation procedure is visualized in the inset. Data belonging to a single patient is always grouped together.

0 derived from a cross-sectional dataset do not directly translate into impact in clinical practice. First, a model with a high AUC and high recall (i.e., sensitivity) can still have poor precision, leading to many false alarms and potentially alarm fatigue. Secondly, the intended use of the early warning system is to provide sepsis risk scores at regular intervals, e.g., hourly. This longitudinal performance of the model is not properly captured when constructing a model on a cross-sectional dataset, even when using different time horizons, since in the longitudinal data the prevalence of LOS will be lower due to repeated measurements and subsequent predictions being correlated. Therefore, a more thorough impact analysis was performed, mimicking the intended clinical use to gain insights into the model performance (i.e., precision, recall and time gained) given a particular configuration of the early warning system.

A simulation was performed to assess the clinical usefulness, henceforth called the *alarm analysis*. The classifiers developed on the cross-sectional dataset were applied to the full longitudinal dataset, generating hourly risk scores for each patient between 72 h after birth until discharge or death. Again grouped cross-validation was used on the training set to ensure no predictions were made for a patient that was used in model training. First, using these predictions AUCs were calculated for different time from 24 h before ($t = -24$) to 12 h after ($t = 12$) the blood culture to gauge early warning performance. For this computation each datapoint in the cross-sectional analysis was replaced by the corresponding point between 24 h earlier and 12 h later. Second, a situation was simulated in which for each patient a prediction is made on an hourly basis starting from 72 h after birth. A clinician is alarmed as soon a risk score crosses a particular threshold. Any alarm between 24 h before and 12 h after a positive blood culture was considered true positive. Any alarm outside this time window was considered a false positive, while the absence of alarms within this time window was considered a false negative. Consecutive predictions are correlated, and clinicians are not expected to return to the same patient if an alarm was raised the previous hour. Therefore, alarms were clustered, and a new alarm was silenced if it fell within a specific time interval, i.e., a refractory period, from the previous alarm. By default, this time interval was set to 8 h, representing the length of a shift at the WKZ NICU (Fig. 3). This implies that a stable but high-risk patient might only have one alarm at the start of the high-risk period.

The thresholds were set to yield either three alarms per day, two alarms per day, one alarm per day, or one alarm per week for the entire NICU. Three alarms per day roughly correspond to the frequency with which the NICU team at the WKZ NICU considers sepsis during their routine. In addition, we simulated a scenario with three thresholds

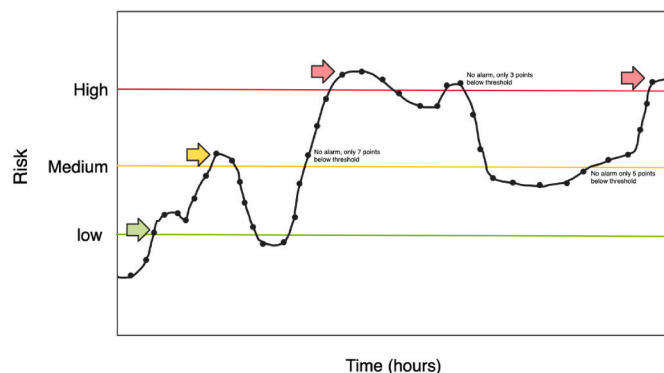


Fig. 3. Example of alarm muting. In the alarm analysis, the full patient time series with hourly predictions are analysed. Alarms are produced when a particular threshold is exceeded (e.g. here at low, medium or high). To prevent alarm fatigue alarms are muted during an 8 h refractory period unless a higher threshold than the one corresponding to the previous alarm has been exceeded. If the alarm level remains high this period extends. In this example there are four alarms, one at each of the arrows.

corresponding to different risk levels. In this scenario, an alarm was also raised during the refractory period if it corresponded to a higher risk level. The three thresholds corresponded to those yielding two alarms per day, one alarm per day, and one alarm per week when used in isolation. With these simulations, the recall, precision, and time gained could be computed. By default, these metrics were computed by only looking at the first positive blood culture for each patient. Longitudinal information of individual patients was grouped and 1000-fold bootstrapping was used to calculate 95% confidence intervals (CIs).

To facilitate clinicians in understanding the patient journey the longitudinal risk scores for each patient were visualized in a heatmap, where each row represents the timeline of a single patient ranging from 72 h after birth until death or discharge. The risk of sepsis over time can be visualized based on a colour scheme ranging from blue (low probability) to red (high probability). These heatmaps helped to identify interesting LOS and control cases, which warranted further investigation.

2.6. Qualitative analysis of patient journey

Finally, an interactive colour-coded dashboard was developed showing hourly risk scores. The colour scale goes from green (low risk) to red (high risk) (e.g., Fig. 3). This dashboard is representative of an actual clinical implementation and together with the results from the alarm analysis was used to iteratively improve the model (Fig. 2).

2.7. Software

All data were extracted from the EMR system MetaVision® (iMDsoft, Tel Aviv, Israel). Pre-processing and modelling were performed in Python 3.8.2 (Python Software Foundation, Beaverton, USA) using the packages Cython 0.29.27, dash 2.0.0, interpretML 0.2.7 [17], matplotlib 3.5.1, NumPy 1.21.5, pandas 1.3.5, scikit-learn 1.0.2, SciPy 1.7.1, SHAP and xgboost 1.0.2 [14]. Statistical analysis of baseline characteristics was performed in SPSS® versions 26 (IBM, New York, United States).

3. Results

3.1. Baseline characteristics

A total of 2519 infants met the inclusion criteria and entered the study, of which 389 infants experienced LOS with a positive blood culture. The remaining 2130 infants were assigned to the control group (Fig. 1). Based on the gestational age distributions of the LOS patients we could not match all control patients. As a result, 1941 control patients were matched (1497 in the train set and 444 in the test set). We did, however, present baseline characteristics for the total patient groups (LOS $n = 389$; control $n = 2130$), except for the age of onset and laboratory data, as equivalent blood culture timings were only available for the matched controls.

Baseline characteristics for LOS patients and control patients, as total groups and separated by train/test label, are summarized in Table 1. The median gestational age of LOS patients was 28.1 weeks and 30.3 in the control group ($p < 0.001$). Also, the median birth weight of both groups was significantly different, with 1020 g in the LOS group and 1300 g in the control group. Patient groups did not differ in terms of sex. The median number of hospitalization days was higher in the LOS group (41 days) than the control group (13 days), not corrected for pregnancy duration and birth weight. Age of onset and laboratory data was compared between LOS patients and matched control patients ($n = 1968$). The median age of onset of LOS, measured as the distance between birth and positive blood culture, was 9 days after birth. In the case of matched control patients, the median artificial age of onset was 4 days after birth. As this is an artificial timestamp, clinical relevance is absent. The artificial difference (9 vs 4 days) is a result of our sampling technique as we took all possible control matches for LOS patients into

Table 1
Baseline characteristics.

	LOS	Control	P value
	Total (n = 389)	Total (n = 2130)	
Demographics			
Gestational age (weeks), median (IQR)	28.1 (26.3, 30.1)	30.3 (28.4, 31.4)	$p < 0.001$
Birth weight (grams), median (IQR)	1020.0 (800.0, 1303.8)	1300.0 (1035.0, 1580.0)	$p < 0.001$
Sex (female), n (%)	191 (49.1)	983 (46.2)	0.31
Hospitalization (days), median (IQR)	41.0 (18.2, 61.2)	13.2 (7.6, 30.7)	$p < 0.001$
Age of onset 1st LOS episode (days), median (IQR) ^b	8.8 (6.6, 12.2)	3.6 (3.5, 5.4)	$p < 0.001$
Laboratory data			
CRP (mg/L), median (IQR) ^{a, b}	18.0 (6.0, 45.0)	2.0 (0.8, 3.8)	$p < 0.001$
CRP >10 mg/L, n (%) ^{a, b}	248 (63.8)	164 (8.5)	$p < 0.001$
Leukocytes (10 ⁹ /L), median (IQR) ^{a, b}	13.1 (7.9, 20.3)	8.8 (6.5, 12.9)	$p < 0.001$
Neutrophils (10 ⁹ /L), median (IQR) ^{a, b}	7.7 (4.3, 13.5)	3.4 (2.1, 5.9)	$p < 0.001$
Thrombocytes (10 ⁹ /L), median (IQR) ^{a, b}	189.7 (111.0, 284.3)	212.4 (153.4, 273.9)	$p < 0.05$
Lines in situ			
Central venous line, n (%) ^c	37 (9.5)	53 (2.5)	$p < 0.001$
Umbilical line (A), n (%) ^c	181 (46.5)	681 (32.0)	$p < 0.001$
Umbilical line (V), n (%) ^c	352 (90.5)	1730 (81.2)	$p < 0.001$
PICC line, n (%) ^c	192 (49.4)	611 (28.7)	$p < 0.001$

^a specimen date nearest to blood culture.

^b data represents only matched control patients n = 1968 (1511 in train and 457 in test).

^c presence of lines during entire hospital stay.

account instead of just one. Control patients have a much shorter hospital stay than LOS patients, but they constitute most of the patient population, explaining the observed difference in onset times. Blood specimen times were used as a proxy for sepsis onset for LOS patients. To reflect on the most relevant laboratory data, the onset times of both LOS and control patients were matched to the closest blood specimen times for other laboratory measurements available in the dataset. All these specimens were drawn approximately 2 h before the sepsis onset time. It remains difficult however to conclude anything based on these findings because CRP levels are routinely checked. Often routine measurements take place around 8 a.m. while many blood cultures for LOS evaluation are taken around 12 a.m., meaning that many CRP measurements precede by default (see supplementary material, Figs. S1 and S2). CRP concentrations in LOS patients were considerably higher than in control patients, with a median of 18 mg/L and 2 mg/L, respectively. Also, leukocyte, neutrophil and thrombocyte counts were significantly higher in LOS patients compared to control patients. These observations indicate the greater inflammatory response of LOS patients. Significantly higher percentages of LOS patients were treated with central venous, PICC, and umbilical lines, which could suggest LOS patients experienced a higher need for medical support. However, it should be noted that these analyses do not include a correction for potential confounders such as gestational age. Post-hoc analysis did reveal that for instance the observed rates between umbilical lines disappears when correcting for gestational age. The similarity between the baseline characteristics of patients in the train and test sets is illustrated in the supplementary material (see supplementary material, Table S2). There were no significant differences observed between any train- or test dataset characteristics, except for the LOS onset in the matched control patients.

3.2. Variable selection

A range of routinely collected monitoring data was evaluated as input data for the algorithm. After careful scrutiny of the data generation process, most were not selected due to having clinician-initiated bias, such as CRP and blood pressure data, which are only measured if there is already a clinical suspicion. Other data such as temperature and respiratory rate are influenced by external factors such as incubator temperature and ventilatory rhythm, introducing artifacts and noise and effectively making them unsuitable. Furthermore, central line and nutrition data showed some potential but were not investigated further. Eventually, this resulted in an algorithm that was built using only heart rate and oxygen saturation measurements, both of which are continuously monitored.

3.3. Algorithm performance

All three classifiers, LR, GAMs and XGBoost, had a comparable performance at $t_{onset} = 0$, with an AUC of 0.73 and 0.79 on the train and test set respectively for the LR model (Fig. 4). The AUC scores for the other models were 0.72 and 0.77 on train and test set for GAMs, and 0.73 and 0.79 for XGBoost respectively (see supplementary material, Figs. S3 and S5). According to the DeLong tests, no significant differences in performance were found between models. These scores indicate that the algorithm is generally able to distinguish patients with sepsis from those without. However, $t_{onset} = 0$ is of limited diagnostic value as at this time the clinician has ordered a blood culture. Since the LR model performs on par with the other models, it will henceforth be used as the default model for which results are presented, unless specified otherwise. Model decisions of the LR model are transparent and sensible from a clinical perspective (see Table 2). Feature importances of the GAM and XGBoost model were in agreement with the LR model (supplementary material Sect. 2.2).

The AUC at different time horizons, from 24 h before ($t = -24$) to 12 h after ($t = 12$) blood culture, provides an indication of algorithm performance as an early warning system. As expected, there is an increase in AUC leading up to the blood culture and a decrease after, e.g., the start of treatment (see Fig. 5). Results for XGBoost and GAMs classifiers are shown in the supplementary material (see Figure SFig. 4-S7). Further analyses for the performance on different age groups, time of birth, gender, and the effect of changing the random seed determining the train/test split are discussed in supplementary material Sect 2.2.2 and Sect 2.2.3.

3.4. Impact assessment

Longitudinal risk profiles for each patient on the cross-validated training set and test set are presented in Fig. 6. High-risk scores are often observed around the time of a positive blood culture. Low-risk control patients typically have a shorter stay on the NICU, whereas high-risk control patients remain on the NICU for extended periods. This might indicate that even though no positive blood culture was taken, other clinical events might have occurred that are associated with an unstable health status. Similar patterns are observed in the test set (lower panel).

The alarm analysis quantifies the observations in the above-mentioned longitudinal heatmaps. We found a total number of days of data between admission of the first until discharge of the last patient of 4025 days with an average number of 17.37 patients on the ward. For the alarm threshold resulting in 2 alarms per day, 80.8% (53.7%) of the LOS (control) patients had at least one alarm during hospitalization. As expected, fewer alarms per day, i.e., higher risk alarms, resulted in a higher precision but lower recall (Table 3). The simulation applying multiple alarm thresholds yields the highest overall recall and an improvement in precision compared to the 2-alarms-per-day scenario. The increase in recall is explained by the possibility to have another

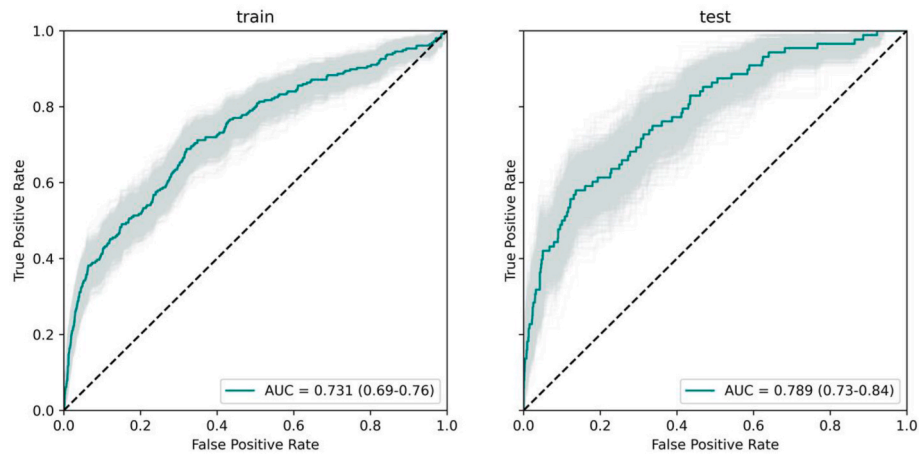


Fig. 4. ROC curves on the cross-sectional dataset for the logistic regression model. The left (right) figure shows the performance of the logistic regression model on the cross-sectional train (test) dataset with LOS patients and matched controls. Errors correspond to 95% C.I.

Table 2

Inputs and coefficients of the logistic regression classifier. The best fit logistic-regression classifier is $p(x) = \frac{1}{1 + e^{-z(x)}}$, with $z(x) = \sum_i \beta_i \cdot g(x_i)$ and $g(x_i)$ a scaling function. For the mean and variance of the heart frequency and the variance of the oxygen saturation, a robust scaler was used.

Variable (x_i)	Feature ($g(x_i)$)	Coefficients (β_i)
Bias	1	-3.99
HF mean (4 h)	$x - 162.5$	0.44
HF variance (4 h)	$\frac{14.5}{x - 75.4}$	0.49
SpO2 variance (4 h)	$\frac{74.1}{x - 4.4}$	0.09
SpO2 min (4 h)	$\frac{9.5}{1 - \frac{x}{100\%}}$	1.62
Bradycardia (4 h)	x	0.10
Tachycardia (4 h)	x	-0.11
SpO2 drops (4 h)	x	0.03

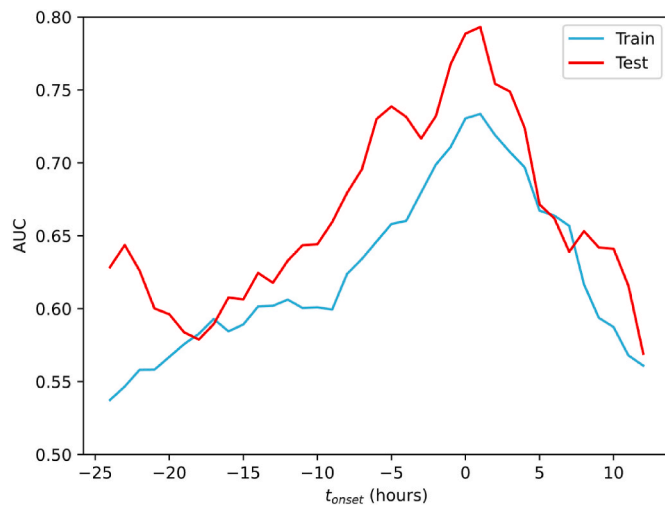


Fig. 5. AUC for different time horizons. AUC of the logistic regression model as a function of t_{onset} for the train (blue) and test (red) set. As expected, performance peaks at the time of the blood culture, $t_{\text{onset}} = 0$.

alarm during the refractory period if the risk threshold is higher than that of the previous alarm. In case the first alarm was outside of the window $(-24, 12)$ hours around the blood culture, the second alarm could still correspond to the LOS episode. For the multiple alarm

analysis of the LR model we find a precision of 3.93%, and recall of 58.6%, clearly outperforming randomly generated alarms (Table 3).

The difference in time between alarms and positive blood cultures provides insight into the suitability of the algorithm as an early warning system. As expected, alarms for LOS patients peak near $t_{\text{onset}} = 0$ and are somewhat offset towards the left, i.e., earlier times (see Fig. 7). This implies that the model has the potential to be on average slightly faster in detecting LOS than the clinician which is, of course, beneficial for an early warning system. For control patients, we use the artificial t_{onset} from the matching procedure. Results are more uniform, although a slight peak can still be observed, an artifact introduced by selecting the nearest alarm to the artificial blood culture moment. Fig. 8 shows the cumulative recall over time for an alarm threshold corresponding to the multiple alarm analysis. Again, for each patient the alarm nearest to the moment a blood culture was taken is shown. This figure allows us to determine what fraction of patients are identified in a certain time window before the blood culture was taken, thereby providing insight into the suitability of the algorithm as an early warning system. The two dashed horizontal lines indicate the recall at $t_{\text{onset}} = -24$ and $t_{\text{onset}} = 0$ (moment of the blood culture) for the training (blue) and test set (red). In the training set, the recall increases from 4.5% to 51.4%, while for the test set this is even a bit higher (3.2%–59.6%). This indicates that the model still reaches a decent recall in the period prior to a blood culture, meaning that a proportion of almost two-thirds of all patients with LOS at least was detected by the model before clinical suspicion.

In Section 2.3 of the supplementary materials, we illustrate how results depend on the configuration of the alarm analysis. Such configurations include other definitions of the time window which define a true positive alarm (Supplementary Tables S3 and S4), whether to analyse only the first or all blood cultures (Supplementary Tables S5 and S6) and other refractory periods (Supplementary Tables S7 and S8).

3.5. Interactive dashboard

An interactive dashboard was developed representative of what an actual clinically implemented early-warning system could look like (Fig. 9). It was used to qualitatively study model performance. Predictions, i.e. risk scores, from the model are shown on the vertical axis from low (green) to high (red). Note that model scores have an arbitrary meaning and cannot be interpreted as a probability since they were derived from the cross-sectional analysis. Therefore, and for an improved user experience, the vertical axis was rescaled. Now, the transitions between colours can be roughly interpreted as 95%, 75%, 50%, 25% and 5% recall, e.g. half the patients experiencing LOS will have at least one risk score in the upper half of the plot in the window

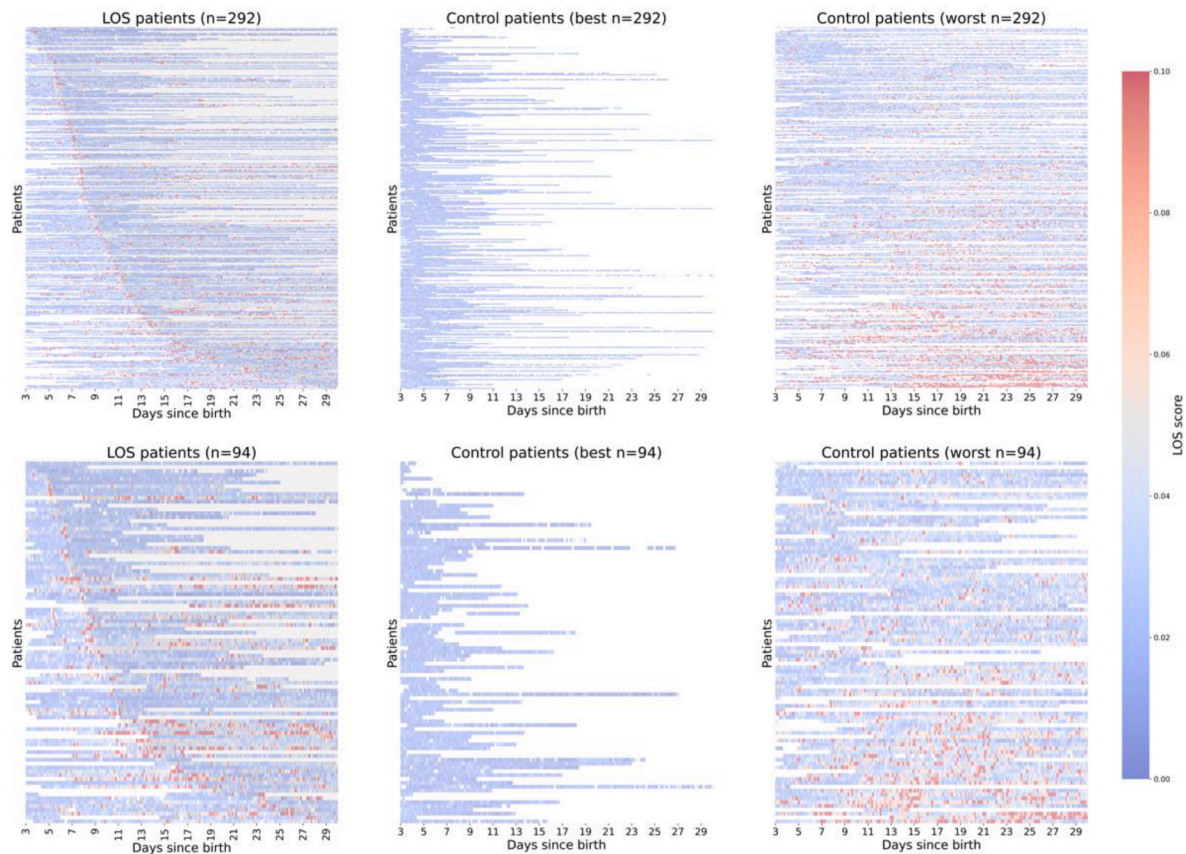


Fig. 6. Patient longitudinal risk scores. Heatmap with the longitudinal risk scores for different patients (rows) from the cross-validated predictions on the training data. Left: LOS patients are ordered by the time of the first positive blood culture (transition to the grey-shaded area). High-risk scores are typically observed around the time of the blood culture. Centre: lowest risk control patients. On average these have a shorter length of stay. Right: highest risk control patients ordered from top to bottom by increasing risk. The upper panel shows the results for the train set and the lower panel for the test set.

from $t_{\text{onset}} = -24$ until $t_{\text{onset}} = 12$. However, this interpretation does not take into account the refractory period used in the alarm analysis, e.g. a patient at a constant high risk for a long time might have had an alarm before $t_{\text{onset}} = -24$, therefore the y-axis does not directly translate to the recall derived from the alarm analysis as shown in Table 3. For illustrative purposes, the thresholds corresponding to three, two and one alarm per day and one alarm per week are also shown. The multi-alarm analysis used all these thresholds except for the three per day. In a real-life implementation, one data point would be added to the dashboard every hour on the right side of the graph. In addition to the risk score, the dashboard also showed patient demographics such as age and gender. Moreover, by clicking on an individual prediction a model explanation is provided through features values and model coefficients (not shown in Fig. 9). For every feature, it can be determined how much it contributes to a particular prediction.

4. Discussion

The present study presented a machine learning algorithm for the early detection of late-onset sepsis in preterm infants. Our model only uses low-frequency heart rate and oxygen saturation data, which are routinely measured at the NICU. Moreover, the logistic regression model is interpretable, the features used are clinically relevant and the model coefficients sensible. An alarm simulation demonstrating performance in clinical practice indicated that about 47% of patients can potentially be identified earlier. Although the false-alarm rate is high, the number of alarms that initiate an action from the clinician is comparable to the number of sepsis checks currently performed at the NICU of the WKZ. For future clinical implementation, risk scores and model explanations

can be efficiently provided in an interactive dashboard.

To the best of our knowledge, the present study presents the most extensive retrospective assessment of clinical applicability to date on the early prediction of LOS. In particular, the quantitative alarm analysis insights into the expected false alarm rate and potential time gained. However, direct comparison with other studies is difficult due to differences in the datasets, study design, and definitions of sepsis. For instance, here only culture-positive sepsis cases were considered. Consequently, some clinical sepsis cases might have been wrongfully classified as either true negative or false positive. Griffin et al. [18] use a combination of heart rate characteristics (HRCs) and laboratory tests to predict LOS, achieving an AUC of 0.82. With HRCs only, which they note are most relevant since laboratory tests are likely clinician-initiated upon sepsis suspicion, they find an AUC of 0.73, comparable to what is presented here. The HeRO² monitor uses HRCs as an early indicator of inflammatory diseases such as sepsis and NEC. Data from a randomized controlled trial³ showed a decreased sepsis-related mortality rate in patients exposed to monitoring, but also an increase in the antibiotic treatment [19,20]. Contrary to the present study, HRCs require high-frequency ECG data making it less readily available [21]. RALIS is another early-warning system for LOS and NEC that corrects the inter-patient variability by using features adjusted for GA and weight [2,22]. In a retrospective case-control study, they achieved an AUC of 0.90. However, the authors consider an alert 7 days before and after a LOS episode a true positive. This wide acceptance range likely gives an

² <https://www.heroscore.com>.

³ ClinicalTrials.gov; NCT00307333.

Table 3

Alarm analysis metrics. Results are shown for the three different classifiers and for a simulation with randomly initiated alarms. Best performance is achieved with a multi-threshold configuration in which alarms can be unmuted by higher threshold alarms. In practice, clinicians at the WKZ perform a sepsis check about three times per day.

	Threshold set for	Train			Test		
		Average number of alarms (n)	Recall (%)	Precision (%)	Average number of alarms (n)	Recall (%)	Precision (%)
Logistic Regression (LR)	1 alarm per week	0.14	16.44 (12.22–20.72)	11.66 (8.55–15.12)	0.16	20.21 (11.93–27.73)	13.92 (8.09–19.69)
	1 alarm per day	1.02	47.95 (42.48–53.69)	5.12 (4.35–6.01)	1.01	51.06 (40.93–61.09)	5.22 (3.85–6.67)
	2 alarms per day	2	56.51 (51.03–62.24)	3.07 (2.63–3.54)	1.93	62.77 (52.80–72.34)	3.40 (2.59–4.23)
	3 alarm per week	3	66.78 (61.14–72.04)	2.47 (2.13–2.80)	2.86	61.70 (51.49–71.14)	2.40 (1.80–3.05)
	Multi-alarm	2.79	58.56 (53.23–64.33)	3.93 (3.38–4.55)	2.75	67.02 (56.98–76.25)	4.41 (3.31–5.51)
General Additive Models (GAMs)	1 alarm per week	0.14	11.64 (8.25–15.26)	8.47 (5.95–11.33)	0.22	25.53 (16.48–34.07)	12.16 (7.53–16.71)
	1 alarm per day	1	43.15 (37.68–48.49)	4.69 (3.95–5.59)	1.16	54.26 (43.35–63.90)	4.70 (3.44–5.94)
	2 alarms per day	2.05	58.22 (52.44–63.84)	3.19 (2.73–3.65)	2.14	60.64 (50.57–70.19)	3.15 (2.35–3.98)
	3 alarm per week	3.06	64.73 (59.04–70.12)	2.41 (2.08–2.76)	3.17	63.83 (53.80–73.19)	2.35 (1.78–2.93)
	Multi-alarm	2.7	59.93 (54.02–65.34)	3.68 (3.18–4.27)	2.89	64.89 (54.73–74.63)	4.16 (3.12–5.18)
eXtreme Gradient Boosting (XGBoost)	1 alarm per week	0.13	13.36 (9.56–17.90)	9.85 (7.11–12.86)	0.16	19.15 (11.83–26.97)	12.50 (7.44–18.07)
	1 alarm per day	0.99	42.47 (37.30–48.23)	4.60 (3.85–5.47)	0.97	51.06 (40.65–61.02)	6.06 (4.39–7.80)
	2 alarms per day	1.96	58.56 (52.72–64.64)	3.39 (2.91–3.93)	1.96	63.83 (53.76–73.48)	3.70 (2.75–4.60)
	3 alarm per week	3.02	68.15 (62.68–73.41)	2.61 (2.27–2.99)	2.78	69.15 (59.88–78.21)	2.89 (2.16–3.65)
	Multi-alarm	2.52	59.25 (53.46–65.33)	4.00 (3.39–4.70)	2.52	65.96 (55.89–75.28)	4.69 (3.53–5.82)
Random alarms (LR)	1 alarm per week	0.14	1.37 (0.32–2.83)	0.93 (0.21–1.91)	0.00 (0.00–0.00)	0.00 (0.00–0.00)	0.00 (0.00–0.00)
	1 alarm per day	1.02	7.53 (4.61–10.84)	0.78 (0.47–1.12)	6.38 (2.09–11.66)	0.59 (0.19–1.11)	6.38 (2.09–11.66)
	2 alarms per day	2	16.44 (12.39–21.08)	0.95 (0.68–1.23)	13.83 (7.10–21.14)	0.72 (0.36–1.18)	13.83 (7.10–21.14)
	3 alarm per week	3	22.95 (18.49–27.97)	0.98 (0.76–1.23)	18.09 (10.23–26.23)	0.90 (0.48–1.40)	18.09 (10.23–26.23)
	Multi-alarm	2.79	16.44 (12.39–21.08)	0.89 (0.63–1.16)	13.83 (7.10–21.14)	0.69 (0.33–1.13)	13.83 (7.10–21.14)

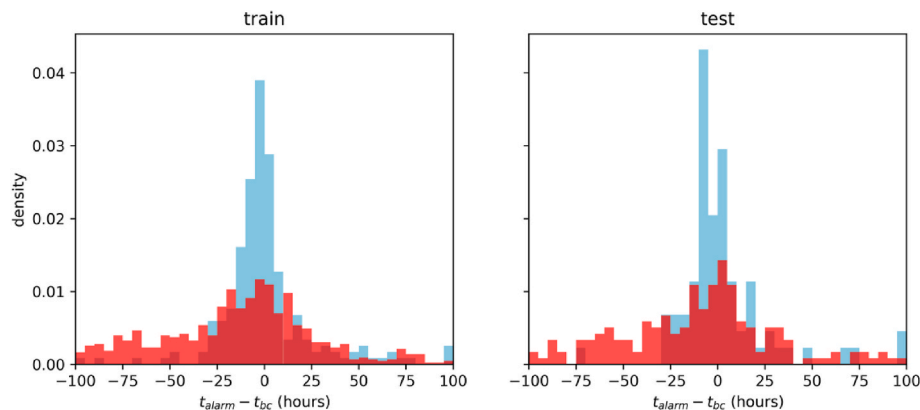


Fig. 7. Distribution of offset between patients' nearest alarm and blood culture. Density plots for the difference between the nearest alarm and time of the blood culture for the train (left) and test (right) set. LOS (matched control) patients are shown in blue (red). All results are for the logistic regression model and a threshold yielding one alarm per day.

overestimation of the real model performance, as the first signs of LOS usually do not start as early as 7 days before clinical suspicion, but false positives are common. Mani et al. [11] trained a model using data from sepsis suspected patients, divided into three groups: culture-positive

sepsis, culture-negative sepsis, and no sepsis. In the algorithm excluding culture-negative sepsis, they achieved an AUC up to 0.78, sensitivity up to 95%, and specificity up to 47%. Their study population consisted of infants suspected of sepsis, unlike our study which considers

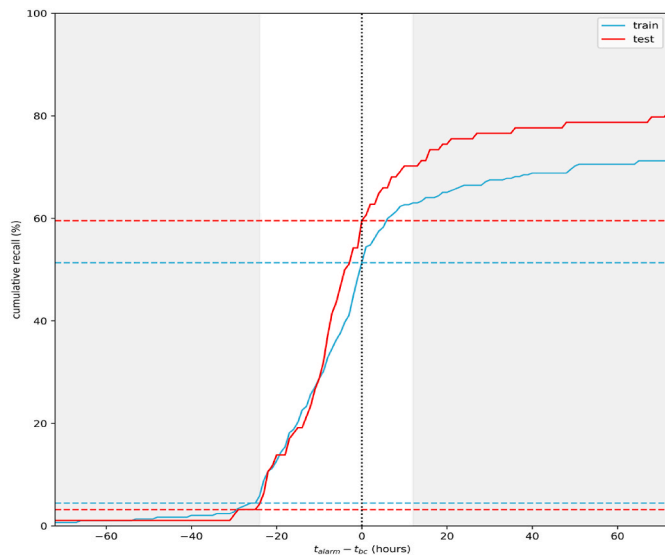


Fig. 8. Cumulative recall over time. The plot displays the cumulative recall over time for both the train (blue) and test (red) data (top panel). For each patient, the nearest alarm to the blood culture from the logistic regression model with the multi-threshold analysis was selected. Dashed horizontal lines indicate the recall between $t_{\text{onset}} = -24$ and $t_{\text{onset}} = 0$, the vertical difference between these lines represents the fraction of patients that can potentially be identified earlier.

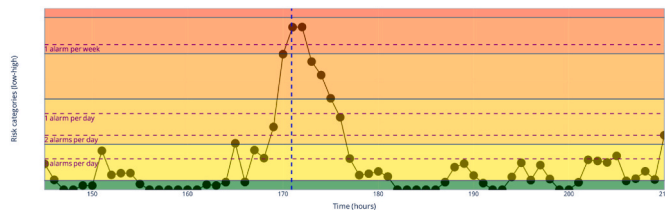


Fig. 9. Visualisation of the dashboard. Risk scores for a patient are shown between 145 and 210 h after birth. In the multi-alarm setup, three alarms would have been raised before the positive blood culture (blue-dashed line), after crossing the two-alarms per day threshold, the one-alarm per day threshold, and after the one alarm per week threshold respectively. After the blood culture antibiotics were administered and the risk score drops.

all patients with $GA \leq 32$ weeks independent of the reason for admission. Moreover, they use a much broader set of features including many laboratory values. Consequently, rather than an early-warning system, their model is better suited to address the question of whether a suspected infant needs antibiotic treatment or not. Similarly, Masino et al. [8] identified three patient groups: culture-positive sepsis, clinical sepsis, and no sepsis. Considering only culture-positive cases, the team achieved an AUC of 0.83 at $t = -4$ h using a wide variety of features. Finally, Cabrera et al. [12] developed a ML model using high-resolution ECG, respiratory and body motion features. They had a mean AUC of 0.79 3 h before the onset of sepsis in a case-control study. Their slightly higher performance in the case-control study could be due to the higher resolution data.

Predictive algorithms for sepsis and septic shock in adult patients have also been an active research field [23]. A small number of algorithms such as the proprietary Insight algorithm have been tested in the practice [24,25]. However, Fleuren et al. [23] find that also here the overall number of clinically implemented algorithms is limited and the results are not yet convincing. The (deep) Aise algorithm is another promising avenue toward a clinically relevant decision support system [26,27]. Indeed, the authors also acknowledge the limited applicability of the AUC when it comes to studying relevant clinical variables such as

the false-alarm rate [27].

As mentioned, a major strength of our approach is in the extensive simulation of clinical practice through the alarm analysis, quantifying both the false-alarm rate and potential time gained. We showed that a NICU protocol ordering a bedside visit at each alarm would not exceed the number of sepsis checks done currently but could potentially result in early detection. Although other works have obtained slightly higher AUCs or used more advanced models, few have attempted to quantify the potential clinical impact in detail. We argue that directly studying the metrics of interest is of utmost importance because when creating a medical device, be it investigational or not, intended use should be specified in conjunction with an adequate performance claim.^{4,5} This performance claim should be suitable to identify any hazard related to using the device in practice, such as over or undertreatment. In addition, we have created a model based on readily available low-frequency monitor data of all patients, allowing for this method to be used in NICUs around the globe. Finally, the model and interface were a co-creation between data scientists and clinicians, resulting in an interpretable model and useable interface.

Several limitations can also be identified. First, we suffered from missing data such as blood culture timestamps because of using all clinical data. The imputation of blood culture timestamps can reduce model performance, as LOS symptoms can present in within a time span of hours. However, we verified that removing blood cultures with contaminated time data did not improve results. Secondly, only clinical sepsis cases were considered positive. Including culture-negative ‘clinical’ sepsis cases could result in better model performance since these patients are likely to be classified as false positives now. Thirdly, the moment of blood culture was used as a proxy for the moment of clinical suspicion. However, since the actual clinical suspicion could come earlier this can lead to an overestimation of the performance of our early warning system. Since other information about clinical suspicion was unavailable further research is needed to assess clinical performance. suNext, our algorithm bases its sepsis prediction on the deterioration of the physiological state. Therefore, it might not be specific to sepsis but could also pick up on other inflammatory diseases such as NEC. Finally, model performance could possibly be improved by using either higher-frequency data or constructing more elaborate features.

The next step in determining the clinical usefulness of the algorithm is a prospective validation study. In a future study, data will be collected not only on the gold-standard positive blood cultures but also on other diseases such as NEC and the clinician’s first sepsis suspicion. This information can be used to assess whether the model can truly provide an early warning. Moreover, if false positive model predictions would nevertheless correspond to suspicion by a clinician without a model, this would be favourable in terms of alarm fatigue, as there is added burden in that case. After prospective validation, the model should be tested for performance in a clinical validation study. Counterfactuals, i.e., would sepsis have been detected later without the early-warning model, are not measured. Therefore, the metrics of interest include indirect performance measures such as reduced mortality, length-of-stay and antibiotics use, similar to the clinical trial of the HRC model [19].

5. Conclusion

The severity and high risk for fulminant sepsis in preterm and VLBW infants, as well as the increasing number of drug-resistant microbes in the NICUs around the world, ask for appropriate surveillance and targeted treatment. In our study, we have built a machine learning algorithm capable of early detecting late-onset sepsis in preterm infants from routinely gathered heart rate and oxygen saturation data. Based on the algorithm’s useable precision and recall, and its ability to detect LOS

⁴ MDR: <http://data.europa.eu/eli/reg/2017/745/oj>.

⁵ IVDR: <http://data.europa.eu/eli/reg/2017/746/2022-01-28>.

earlier with respect to the clinician, implementation in clinical practice seems to be promising.

In conclusion, we have captured early clinical signs resulting from LOS to predict LOS before clinical suspicion arises. We have built a working algorithm (NICU EW), potentially clinically relevant, from routinely gathered data. Its performance has been validated in a cross-sectional dataset as well as in a more realistic longitudinal simulation. The next step would be to evaluate it in a prospective study and accurately determine the real-time clinical value. If so, NICU EW may soon help speed up clinical decision-making for treatment, alongside doctor's eye and laboratory assessment.

Author contributions

DV and MB initiated the Big Data for Small Babies (BD4SB) study and made the study design. Data analysis was performed by MvdB, OM, IL and RB. MvdB, OM, RB, and DV wrote the manuscript. MvdF, JD and MB scrutinized the analysis for clinical relevance. All authors provided feedback on the manuscript and read and approved the final version for submission.

Data and materials availability

The analysis code and de-identified data underlying the results reported in this article can be made available as part of further research collaborations via UMCU's digital research environment (<https://www.andrea-cloud.eu>). Interested parties should contact the corresponding author (DV). Any data sharing will be subject to meeting the Privacy Regulations of UMC Utrecht, the General Data Protection Regulation (GDPR) and the General Data Protection Regulation Implementation Act.

Declaration of competing interest

The authors have not received funding for writing this manuscript. There are no competing interests related to this manuscript.

Acknowledgments

The authors acknowledge support from the Applied Data Analytics in Medicine (ADAM) program of the UMC Utrecht in the early stages of this project. We thank Annemarie van 't Veen, Saskia Haitjema, Lieke van Schaijk and Ruben Peters for their contributions to this project.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.combiomed.2023.107156>.

References

- [1] S.A. Qazi, B.J. Stoll, Neonatal sepsis: a major global public health challenge, *Pediatr. Infect. Dis. J.* 28 (2009) 2008, <https://doi.org/10.1097/INF.0b013e31819587a9>, -9.
- [2] L.B. Mithal, R. Yogev, H.L. Palac, D. Kaminsky, I. Gur, K.K. Mestan, Vital signs analysis algorithm detects inflammatory response in premature infants with late onset sepsis and necrotizing enterocolitis, *Early Hum. Dev.* 117 (2018) 83–89, <https://doi.org/10.1016/j.earlhumdev.2018.01.008>.
- [3] C.P. Hornik, P. Fort, R.H. Clark, K. Watt, D.K. Benjamin, P.B. Smith, et al., Early and late onset sepsis in very-low-birth-weight infants from a large group of neonatal intensive care units, *Early Hum. Dev.* 88 (2012) S69–S74, [https://doi.org/10.1016/S0378-3782\(12\)70019-1](https://doi.org/10.1016/S0378-3782(12)70019-1).
- [4] B.J. Stoll, N. Hansen, A.A. Fanaroff, L.L. Wright, W.A. Carlo, R.A. Ehrenkranz, et al., Late-onset sepsis in very low birth weight neonates: the experience of the NICHD Neonatal Research Network, *Pediatrics* 110 (2002) 285–291, <https://doi.org/10.1542/peds.110.2.285>.
- [5] C.P. Hornik, P. Fort, R.H. Clark, K. Watt, D.K. Benjamin, P.B. Smith, et al., Early and late onset sepsis in very-low-birth-weight infants from a large group of neonatal intensive care units, *Early Hum. Dev.* 88 (2012) S69–S74, [https://doi.org/10.1016/S0378-3782\(12\)70019-1](https://doi.org/10.1016/S0378-3782(12)70019-1).
- [6] A.J. Masino, M.C. Harris, D. Forsyth, S. Ostapenko, L. Srinivasan, C.P. Bonafide, et al., Machine learning models for early sepsis recognition in the neonatal intensive care unit using readily available electronic health record data, *PLoS One* 14 (2019) 1–23, <https://doi.org/10.1371/journal.pone.0212665>.
- [7] S.K. Korang, S. Safi, C. Nava, G. Greisen, M. Gupta, U. Lausten-Thomsen, et al., Antibiotic regimens for late-onset neonatal sepsis, *Cochrane Database Syst. Rev.* 5 (2021) CD013836, <https://doi.org/10.1002/14651858.CD013836.pub2>.
- [8] A.J. Masino, M.C. Harris, D. Forsyth, S. Ostapenko, L. Srinivasan, C.P. Bonafide, et al., Machine learning models for early sepsis recognition in the neonatal intensive care unit using readily available electronic health record data, *PLoS One* 14 (2019) 1–23, <https://doi.org/10.1371/journal.pone.0212665>.
- [9] V.S. Kuppala, J. Meinzen-Derr, A.L. Morrow, K.R. Schibler, Prolonged initial empirical antibiotic treatment is associated with adverse outcomes in premature infants, *J. Pediatr.* 159 (2011) 720–725, <https://doi.org/10.1016/j.jpeds.2011.05.033>.
- [10] A.C. Helguera-Repetto, M.D. Soto-Ramírez, O. Villavicencio-Carrisoza, S. Yong-Mendoza, A. Yong-Mendoza, M. León-Juárez, et al., Neonatal sepsis diagnosis decision-making based on artificial neural networks, *Front. Pediatr.* 8 (2020), <https://doi.org/10.3389/fped.2020.00525>.
- [11] S. Mani, A. Ozdas, C. Aliferis, H.A. Varol, Q. Chen, R. Carnevale, et al., Medical decision support using machine learning for early detection of late-onset neonatal sepsis, *J. Am. Med. Assoc.* 21 (2014) 326–336, <https://doi.org/10.1136/amiainl-2013-001854>.
- [12] L. Cabrera-Quiros, D. Kommers, M.K. Wolvers, L. Oosterwijk, N. Arents, J. van der Sluijs-Bens, et al., Prediction of late-onset sepsis in preterm infants using monitoring signals and machine learning, *Crit. Care Explor.* 3 (2021), e0302, <https://doi.org/10.1097/cce.0000000000000302>.
- [13] B.K. Beaulieu-Jones, W. Yuan, G.A. Brat, A.L. Beam, G. Weber, M. Ruffin, et al., Machine learning for patient risk stratification: standing on, or looking over, the shoulders of clinicians? *Npj Digit. Med.* 4 (2021) 1–6, <https://doi.org/10.1038/s41746-021-00426-3>.
- [14] T. Chen, C. Guestrin, XGBoost: A Scalable Tree Boosting System, 2016, <https://doi.org/10.1145/2939672.2939785>.
- [15] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, et al. (Eds.), *Advances in Neural Information Processing Systems*, vol. 30, Curran Associates, Inc., 2017.
- [16] E.R. DeLong, D.M. DeLong, D.L. Clarke-Pearson, Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach, *Biometrics* 44 (1988) 837–845.
- [17] H. Nori, S. Jenkins, P. Koch, R. Caruana, InterpretML: A Unified Framework for Machine Learning Interpretability, 2019, pp. 1–8.
- [18] M.P. Griffin, D.E. Lake, J.R. Moorman, Heart rate characteristics and laboratory tests in neonatal sepsis, *Pediatrics* 115 (2005) 937–941, <https://doi.org/10.1542/peds.2004-1393>.
- [19] J.R. Moorman, W.A. Carlo, J. Kattwinkel, R.L. Schelonka, P.J. Porcelli, C. T. Navarrete, et al., Mortality reduction by heart rate characteristic monitoring in very low birth weight neonates: a randomized trial, *J. Pediatr.* 159 (2011) 900–906.e1, <https://doi.org/10.1016/j.jpeds.2011.06.044>.
- [20] K.D. Fairchild, R.L. Schelonka, D.A. Kaufman, W.A. Carlo, J. Kattwinkel, P. J. Porcelli, et al., Septicemia mortality reduction in neonates in a heart rate characteristics monitoring trial, *Pediatr. Res.* 74 (2013) 570–575, <https://doi.org/10.1038/pr.2013.136>.
- [21] M.P. Griffin, T.M. O'Shea, E.A. Bissonette, F.E. Harrell, D.E. Lake, J.R. Moorman, Abnormal heart rate characteristics preceding neonatal sepsis and sepsis-like illness, *Pediatr. Res.* 53 (2003) 920–926, <https://doi.org/10.1203/01.PDR.0000064904.05313.D2>.
- [22] I. Gur, G. Markel, Y. Nave, I. Vainshtein, A. Eisenkraft, A. Riskin, A mathematical algorithm for detection of late-onset sepsis in very-low birth weight infants: a preliminary diagnostic test evaluation, *Indian Pediatr.* 51 (2014) 647–650, <https://doi.org/10.1007/s13312-014-0469-x>.
- [23] L.M. Fleuren, T.L.T. Klausch, C.L. Zwager, L.J. Schoonmade, T. Guo, L. F. Roggeveen, et al., Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy, *Intensive Care Med.* 46 (2020) 383–400, <https://doi.org/10.1007/s00134-019-05872-y>.
- [24] J.S. Calvert, D.A. Price, U.K. Chettipally, C.W. Barton, M.D. Feldman, J.L. Hoffman, et al., A computational approach to early sepsis detection, *Comput. Biol. Med.* 74 (2016) 69–73, <https://doi.org/10.1016/j.combiomed.2016.05.003>.
- [25] D.W. Shimabukuro, C.W. Barton, M.D. Feldman, S.J. Mataraso, R. Das, Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: a randomised clinical trial, *BMJ Open Respir. Res.* 4 (2017), e000234, <https://doi.org/10.1136/bmjresp-2017-000234>.
- [26] S. Nemati, A. Holder, F. Razmi, M.D. Stanley, G.D. Clifford, T.G. Buchman, An interpretable machine learning model for accurate prediction of sepsis in the ICU, *Crit. Care Med.* 46 (2018) 547–553, <https://doi.org/10.1097/CCM.0000000000002936>.
- [27] S.P. Shashikumar, C.S. Josef, A. Sharma, S. Nemati, DeepAISE – an interpretable and recurrent neural survival model for early prediction of sepsis, *Artif. Intell. Med.* 113 (2021), 102036, <https://doi.org/10.1016/j.artmed.2021.102036>.